构建以事件为核心的长期保存系统起源管理框架 *

吴振新[1] 李文燕[2] 蒋世银[1]
[1] 中国科学院文献情报中心 北京 100190 [2] 中新金桥数字科技(北京)有限公司 北京 100096

摘要:[目的/意义]研究建立长期保存系统起源管理框架,通过有效管理起源信息,确保长期保存系统所存档数据的真实可靠可用。[方法/过程]基于数字对象保存周期进行起源事件定义,基于 OAIS 保存流程进行起源管理框架设计,以事件为核心进行起源管理功能模型和起源信息模型设计。[结果/结论]初步完成基于事件的保存系统起源管理框架的设计,既遵循保存领域的相关标准,同时兼顾实践需求,对长期保存系统具有很好的普适性和可行性,但其在有效性和实用性方面还有待进一步验证。

关键词: 起源信息 起源事件 长期保存 生命周期管理 OAIS PREMIS PROV 语义模型
分类号: G250.76
DOI: 10.13266/j.issn.0252-3116.2016.06.014

数字对象的起源信息,即 provenance,记录了数字对象的变化历史。通过起源信息,人们可以全面了解数字对象产生之后所发生的变化以及变化的原因、时间、地点、相关人员等 7W 信息(what、where、who、when、which、why、how)。

数字资源长期保存系统作为一类特殊的数据管理系统,通过摄入、保存、管理等一系列管理行为,确保数字对象经过足够长时间后还能够被目标用户团体所使用,对于数据的真实可靠可用,其面临着更大的责任和挑战。在长期保存系统中,起源信息能够发挥多方面的作用。一方面长期保存系统要在相当长的时间内管理和保持数字对象的可用性,另一方面它要抵抗技术变化对数字对象及保存系统所带来的影响。格式迁移、媒体迁移、技术更新是长期保存的常用策略,所以无论是对象本身发生变化还是其所处环境发生改变,都可以利用起源详细记录这些改变,并维护这些变化前后的数字对象的关联。通过这种方法,保存系统能够有效进行版本和衍生物管理,并为数字对象的真实性和系统的可信赖认证提供证据,同时还能为权限管理和责任归属提供支持。因此,起源对于保存系统有着更为重要的意义。

本文作者曾在《起源技术在长期保存中的研究与应用》[1]一文中全面总结和分析了起源在长期保存中的研究情况,初步提出了一个起源管理框架,本文将基于该文,概述如何进一步完善长期保存系统起源管理框架和相关功能的设计。

1 国内外起源研究分析

国外对数据起源技术的研究较早,兴起于 20 世纪 90 年代,涉及计算机、地理系统、生物、金融等多个领域和学科,可分为基础理论和应用两个方面。

起源的基础研究包括起源的定义、组织模型、描述词汇、序列化形式等。模型是起源研究最大的热点之一,目前通用模型有 W3C PROV-DM、OPM、Provenir 和 CRM_{dig} 等。此外,一些元数据方案、本体词汇等也提供了表达 Provence 的术语,如 DC 元数据、VoID 词汇和 Provence Vocabulary。其中最具有代表性的是 PROV-DM,它在 2013 年 4 月被 W3C 发布为起源标准。该模型可以兼容其他起源模型,得到了各个行业的一致认可,使起源信息在系统之间的交互、传递,尤其是在 WEB 环境中大规模地推广和使用成为可能,推动了起源标准化进程。

起源的应用研究包括起源捕获、存储、查询、可视化的技术、工具、系统和框架等。可支持起源管理的工具或系统非常多,例如 Taverna、VisTrails、REDUX 和 VDS 等,它们都可以捕获、存储和浏览工作流环境中的起源,Y. L. Simmhan 等[2-3]对此作了较为全面的综述。起源框架是另一个研究重点,R. Bose 等[4]提出适应于

作者简介:吴振新(ORCID:0000-0003-4966-1961),研究馆员,E-mail:wuzx@mail.las.ac.cn;李文燕(ORCID:0000-0002-7695-5087),助理工程师,硕士;蒋世银(ORCID:0000-0002-2038-4027),助理馆员,硕士。

收稿日期:2016-02-29 修回日期:2016-03-06 本文起止页码:91-96 本文责任编辑:王善军

科技工作流起源管理的概念框架，它可以从工作流系统中自动捕获起源信息，记录为元数据，保存到相应的数据库中提供查询；B. R. Barkstrom 等[5]设计了地球科学数据起源追踪的计算框架，用于追踪一条地球科学数据的创建、保管历史、知识产权历史 3 种类型的活动，并将其呈现为有向无环图；DERI（Digital Enterprise Research Institute）发布了通用的来源管理框架及系统 Prov4J[6]，使用语义网标准和工具管理起源，帮助用户开发起源感知应用程序[7]。

近年来国内也开始了起源的研究，主要涉及起源技术综述[8-9]、起源表达、起源模型分析等。其中模型的研究较多，除了模型综述和述评外[10]，国内学者尝试修改完善已有模型或者提出新的起源模型[11]，如针对 OPM（Open Provenance Model）的安全性改进[12]、基于 DNA 双螺旋结构的数据起源模型[13]等。但国内开展起源管理实践的相对较少，尤其是过程级起源管理，以及起源在特定领域的研究和应用。

在长期保存领域，OAIS 和 PREMIS 等标准给出了关于起源的概念定义和解释，但缺乏相关实践的内容，如记录的内容、相关技术、起源管理策略等。DAITSS、CASPAR、APASEN 等[1]项目已经开始相关探索，国内也有学者[14-16]开始了起源的捕获研究。总体来说，保存领域对起源还缺乏全面、系统化的分析研究。本文从保存生命周期的角度对起源的内容、捕获、存储和封装进行全面的分析，提出一个综合性的完整起源管理框架，希望为长期保存提供有益的参考。

2　长期保存系统的起源管理框架的基本设计思路

从起源在已有长期保存系统的应用中发现，不同的长期保存系统对起源记录的方式、内容和侧重角度均不相同，包括重要操作、责任人、时间和设备等，同时 OAIS 和 PREMIS 也未对起源描述作出详细规范，但是 PREMIS 框架和长期保存实践逐渐形成了以事件为核心来记录起源的方式，所以本文选择以事件为核心来管理起源，其定义为内容信息的历史，展示了内容信息从产生以后发生的相关变动[17]。

起源信息的管理贯穿数字对象在 OAIS 系统中的整个生命周期。考虑诸多因素，本文在起源管理框架的设计中，遵循如下的设计原则：

（1）基于 OAIS。OAIS 是长期保存的通用标准，提供了长期保存的基本流程和事件，所以它是本文研究框架的最基本的出发点。

（2）基于数字对象的保存周期。本文以数字对象提交到保存系统为起点，对其进入保存系统后的整个保存周期内所有变化来实施起源采集与管理。数字对象提交到保存系统前的起源信息，可由内容生产者在和保存方协商一致的基础上以规范的方式提交到长期保存系统，这一点和 OAIS 的要求也是一致的。

（3）以事件为核心来记录起源信息。在长期保存的过程中，数字对象会因各种管理活动产生多种事件并产生起源信息，可以说事件通常都伴随着起源信息的产生。

（4）交互性。信息模型是系统软件设计中重要的内容，为增强起源信息在不同系统之间的交互性，起源管理应采用信息模型来组织管理数据。

（5）通用性。本框架旨在为长期保存中组织和管理起源信息提供一般的功能流程、模型等内容的参考，与具体技术实现无关。

为了明确说明整个管理框架的设计原理及关键内容，本文首先根据保存周期管理提出了起源事件清单，指明哪些事件会产生起源信息、应该在哪些流程实施管理。进而利用 OAIS 保存系统的功能模型示意图，进一步明确起源管理模块与保存系统功能框架之间的嵌入关系，并在起源信息管理的功能模型部分，明确起源管理功能如何融入 OAIS 管理流程实施有效管理。在最后部分则通过起源信息模型，定义起源信息包含基本组成和结构，以便进行信息组织和保存。这 4 个部分基本讲明了保存系统对于起源管理的 what、when、where、how 几个问题。

3　面向保存周期管理的起源事件清单

事件是对象发生变化的主要驱动，通过事件可以把各种类型的状态变化串联起来。随着事件的累积，发生在对象上的事件链可以动态地呈现保存对象的状态改变。因此本文的起源事件定义为涉及或影响至少一个对象或代理的保存系统可识别的动作，和计算机中常说的单击、双击、窗体加载等事件是不同的，它是长期保存系统定义的保存系统处理对象的操作或操作集，如压缩文件、摄入信息包、创建对象等。识别记录哪些事件作为起源事件是本文所构建的起源管理框架的核心问题。

PREMIS[18]作为保存领域的保存元数据标准，其定义了 5 个基本实体，事件是其中之一，因此使用事件记录起源便于使用长期保存元数据进行描述。PREMIS 中定义了 15 种保存事件：creation、deacession、decom-

pression、decryption、deletion、digital signature validation、dissemination、fixity check、ingestion、message digest calculation、migration、normalization、replication、validation、virus check。但这些事件不是专门针对起源设计的起源事件，同时不同事件之间还有重叠的地方，如 creation 和 normalization；部分事件指向不明，在事件应用中可能会产生歧义，如 validation。所以在实际应用中，需要明确定义长期保存系统中哪些事件应该被记录为起源事件。

OAIS 认为：起源是内容信息的历史，它展示了内容信息产生的由来、从产生以后发生的变化以及自创建以后的保管责任方的改变。这个定义暗含了选择起源事件的两个重要依据，即"时间"和"变化"。"时间"，即数字对象的保存周期，在 OAIS 中整个保存周期包含了 6 个保存流程，即摄入(ingest)、归档存储(archival storage)、数据管理(data Management)、业务管理(administration)、保存规划(preservation planning)和访问(access)。"变化"，即判断一个事件是否为起源事件的依据。

综上所述，在遴选起源事件时应考虑如下方面：①导致内容对象最初产生，这是一个由无到有的过程。②导致内容对象本身发生变化，或者能够捕获长期保存过程中，发生的变化，这些变化涉及内容、结构、数量、格式、位置、元数据和保管责任方等方面。③导致新版本对象的产生，虽然数字对象内容本身并未发生任何变化，这是长期保存之基本要义，但是产生了和该数字对象关系密切的新对象，例如副本，不同格式的新版本，对于对象的复用都很有益处。④导致数字对象的版权和管理权发生变化。⑤导致数字对象的消亡。

根据以上原则，从 OAIS 所包括的流程中遴选出如表 1 所示的起源事件。

表 1 长期保存起源事件清单

事件	英文名称	涉及流程
捕获	capture	摄入
解密	decryption	摄入
逆压缩	decompression	摄入
压缩	compression	摄入
病毒检查	virus check	摄入、归档存储
数量检查	number check	摄入、归档存储
内容检查	content check	摄入、归档存储
不变性检查	fixity check	摄入、归档存储
备份检查	backup check	归档存储
硬件检测	hardware detection	摄入、归档存储
格式检查	format check	摄入、归档存储
目录删除	deaccession	归档存储
生产者登记	producer register	摄入
摄入消息计算	message digest calculation	摄入
规范化	normalization	摄入
当前化	contemporary	摄入、归档存储
元数据抽取	description	摄入
创建	creation	摄入
摄入	ingestion	摄入
备份	replication	摄入、归档存储
媒体迁移	media migration	归档存储
删除	delete	归档存储
数据恢复	data recovery	归档存储
模式更新	schema update	数据管理
分发	dissemination	访问
数据迁移	transfer	访问

4 嵌入 OAIS 保存流程的起源管理框架

从起源事件清单可以看出，起源事件涉及 OAIS 的所有流程和功能模块，起源信息的管理需要嵌入到保存系统的完整流程中，见图 1。

其中，图 1 中心部分是起源的管理模块，它要嵌入 OAIS 的各个流程中，动态地监测各个流程的事件，并根据起源管理中预先配置好的起源事件清单，捕获数字对象的起源事件。然后把长期保存捕获的事件和生产者提供的事件按照相应的起源模型组织成规范的起源，存储为相应的格式(封装与存储)，并由保存管

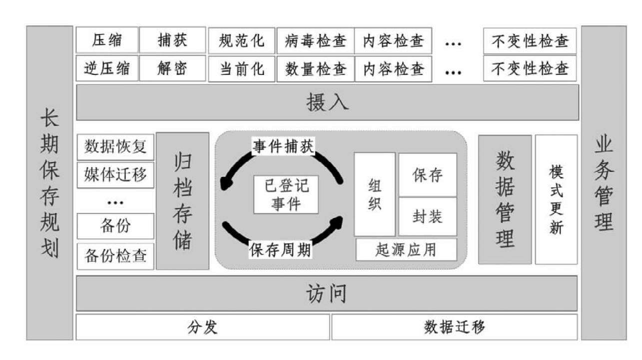


图 1 嵌入 OAIS 保存流程中的起源管理框架

理模块对组织模块生成的起源进行长期保存和管理，保证起源的完整性、可理解性和长期可访问性，同时应用模块按保存系统要求为用户或长期保存的其他模块(如真实性管理)提供起源的使用。

5 以事件为核心的起源管理功能模型

图 2 清晰地展示了起源管理各个功能模块的相互关系及数据流向。该模型包含了 4 个基本子功能模块即捕获、组织、保存管理和应用，其中捕获、组织和保存

管理是功能模型的重点。

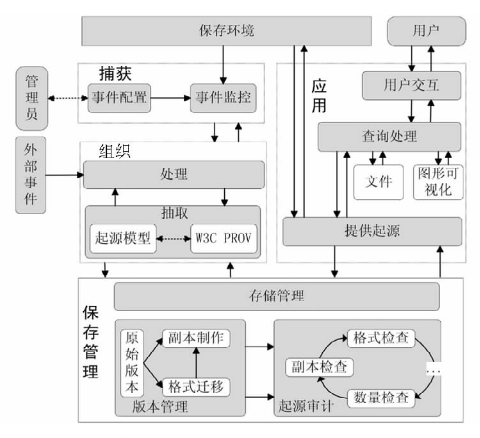


图 2　以事件为核心的起源管理功能模型

5.1　捕获模块

事件配置功能负责预定义和配置起源管理需要捕获的事件类型。这个功能在捕获之前完成，由长期保存系统的管理人员根据保存系统的功能所包含的各个操作，归纳出需要记录为起源的事件，对其进行详细的定义，并把起源事件清单配置为计算机可读的格式，如数据库表或者 XML 文件。

事件监控功能负责动态地监测保存系统所发生的所有事件，当某个事件和预定义起源事件清单中事件相匹配时，则触发组织模块，把事件信息如事件内容、事件时间、操作对象和使用设备等内容传递给组织模块。

5.2　组织模块

组织模块接收到事件消息后将其添加到起源记录的任务队列中，供抽取功能使用。通过这种异步记录的方式来组织起源，既能减小对系统原有进度的影响，又能减少服务器的负担。起源记录的任务队列包括两种类型事件：一种是自动捕获发生在保存统内部的事件，另一种是由被保存内容的生产者提供的外部事件。

抽取功能按照顺序读取任务队列的事件信息，根据系统设定的信息模型（如 XML schema）对事件信息进行规范化组织，生成规范化起源。

5.3　保存管理模块

存储管理功能把从组织模块接收到的起源按照相关方式进行存储，并维护起源数字对象之间的关联关系。版本管理功能实现对起源的各个版本的管理。主要是按照保存计划和政策进行副本制作或者支持格式迁移。起源审计功能负责为每个版本的起源定期执行不变性检查（fixity check）、格式检查和副本检查。触发该功能有两种类型的任务：一种是定期检查任务，一种是新增加起源、备份起源和改变起源版本触发的任务。

5.4　应用模块

应用模块为系统的其他模块（如审计追踪）使用起源信息提供标准的接口调用。查询处理功能直接接收用户的起源请求，然后调用相关的起源接口，返回特定格式的起源给用户。用户交互功能则为用户提供可视化界面，如网页。起源查询或下载的请求消息由用户在交互界面中发起后被传递给查询处理模块，将底层处理后从交互界面返回用户请求的特定格式起源。

6　以事件为核心的起源信息模型

使用信息模型不仅可对所管理的数据进行有效组织，同时也有利于长期保存、管理和重用。OAIS 中数字对象包括内容信息和表征信息两部分，因此起源不但要记录内容信息的变化，也同时需要记录表征信息的变化。起源信息中应包含以下相关内容：

（1）事件。事件的驱动使得数字对象发生变化。事件的细节描述是起源信息的重点内容，除了包含事件标识符、细节描述、时间、事件类型、处理设备、处理结果、地点和发生原因，还要包含事件涉及的责任人和被操作的对象基本信息。

（2）数字对象。需要完整记录事件涉及到的数字对象。通过在事件中引用数字对象的标识符将二者关联起来，但不包含数字对象的描述元数据。如果一个事件同时关联两个甚至更多数字对象，就意味着所有对象都拥有该起源信息，应该包含所有的数字对象的标识符。

（3）代理内容。狭义的代理指的是事件的操作人，此处的代理的含义更广，包括组织、个人、软件和硬件 4 种内容。

（4）数字对象之间的关系。事件对数字对象的操作可能会导致新版本对象产生，如副本或不同格式的版本。虽然数字对象之间的关系可能不被直接记录在起源信息中，但是通过分析相关事件的性质和涉及的输入、输出对象可以间接得出数字对象的版本变化。

W7 语义模型[19]给出了从 7 个维度来记录起源信息的思路，较全面地说明了起源信息包含的内容，对于构建起源模型具有重要的参考作用。本文在 W7 语义模型和上文归纳的事件起源内容基础上设计如图 3 所示的事件起源信息模型，此模型从事件的角度描述了起源信息应该包含的内容元素，并从 7 个维度来设计每个起源事件包含的内容概念。

在该信息模型中，以事件为中心把记录数字对象

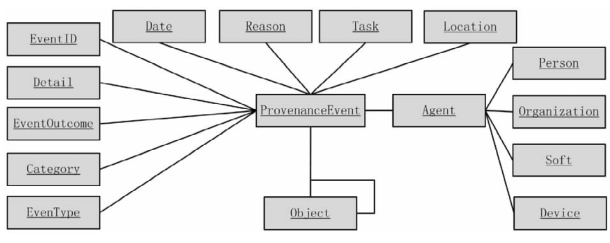


图3 事件起源信息模型

变化的各种信息串联起来,涵盖了起源事件的基本元素:Object、Agent、EventID、Date、Reason、Task、Detail、EventOutcome、Category、EventType 和 Location,每一个元素都能对应表示 W7 模型的一个维度,如表 2 所示。虽然 Agent 和 Object 是事件的一部分,但两者的描述元数据方案不在起源管理框架模型范围内,该模型只对起源事件应该包含的各个概念加以描述。

表2 事件起源信息模型的基本元素

基本元素	解释	对应维度
EventID	事件的唯一标识符,用以引用该事件	what
Date	记录事件发生的时间,可能是一个时间点,也可能是一个时间段	when
Detail	事件的详细内容的文字描述,例如"成功执行对图片的不变性检查"	what
EventOutcome	事件的执行结果描述	what
EventType	事件类型,如病毒检查、不变性检查,推荐使用受控词汇	what
Category	事件分类,如病毒检查和不变必验证均属于验证类事件,对不同事件进行[illegible]方便查询	what
Task	驱动事件发生的任务技术,类似于工作流的定义	how
Reason	驱动事件发生的原因描述	why
Location	和事件相关的位置信息	where
Agent	和事件相关的代理,包括 4 种子类型,Person(人)、Organization(组织)、Soft(软件)和 Device(物理设备)	who,which
Object	事件涉及的被保存对象	what

笔者在系统实现时复用了 PREMIS OWL[20] 和 W3C PROV-O[21] 来实现起源组织模型,并利用 RDF 进行了起源封装;本文没有涉及起源的存储和封装策略,这些内容将在笔者的另一篇论文中详细介绍。

7 结语

目前该框架系统原型刚刚完成,其有效和实用性还有待进一步验证。总地来看,基于事件的保存系统起源管理框架的设计,不但遵循了保存领域的相关标准,同时也参考了许多国际项目实践[1],兼顾了理论和实践两个方面,对长期保存的信息系统具有很好的普适性和可行性。该框架基于数字对象的保存周期,提出了嵌入到 OAIS 流程中与相关保存事件相融合的方法,保证了起源的有效和完整。而以事件来记录起源的方式可以有效地把分散的元数据碎片按照时间组织起来,既紧抓起源产生的本质,又便于起源信息采集管理,这一点和 PREMIS 认为起源以事件为驱动的观点是一致的。另外,本框架的起源信息组织采用了信息模型的设计,与具体的实现技术无关,具有高度的抽象性,可以适合大多数长期保存系统。

保存领域的研究人员已经充分认识到起源对长期保存的重要作用,不断探索如何对起源进行有效管理和使用。希望本文所做的研究和努力,能够为相关人员在长期保存的起源管理理论研究和实践方面提供有益的参考。

参考文献:

[1] 吴振新, 李文燕. 起源技术在长期保存中的应用与研究[J]. 图书情报工作, 2015, 59(8):118-125.
[2] SIMMHAN Y L, PLALE B, GANNON D. A survey of data provenance in e-science[J]. ACM SIGMOD record, 2005, 34(3): 31-36.
[3] SIMMHAN Y L, PLALE B, GANNON D. A survey of data provenance techniques[EB/OL]. [2016-02-16]. http://www.cs.indiana.edu/pub/techreports/TR618.pdf.
[4] BOSE R. A conceptual framework for composing and managing scientific data lineage[EB/OL]. [2016-02-16]. http://ieeexplore.ieee.org/stamp/stamp.jsp? tp=&arnumber=1029701.
[5] BARKSTROM B R. A mathematical framework for earth science data provenance tracing[J]. Earth science informatics, 2010, 3(3):167-196.
[6] Prov4J (Provenance for the Web)[EB/OL]. [2015-05-13]. http://prov4j.org/.
[7] FREITAS A, LEGENDRE A, O'RIAIN S, et al. Prov4J:a semantic Web framework for generic provenance management[EB/OL]. [2016-02-16]. http://people.csail.mit.edu/pcm/tempISWC/workshops/SWPM2010/paper_8.pdf.
[8] 戴超凡,王涛,张鹏程. 数据起源技术发展研究综述[J]. 计算机应用研究,2010(9):3215-3221.
[9] 沈志宏,张晓林. 语义网环境下数据溯源表达模型研究综述[J]. 现代图书情报技术,2011(4):1-8.
[10] 李文燕,吴振新.起源信息模型及标准 PROV 的研究分析[J]. 情报理论与实践,2015,38(4):23-29.
[11] 陈颖. 一种基于 DNA 双螺旋结构的数据起源模型[J]. 现代图书情报技术,2008(10):11-15.
[12] 刘通. 基于 OPM 的安全起源研究[D]. 淄博:山东理工大学,2013.
[13] 陈颖. 一种基于 DNA 双螺旋结构的数据起源模型[J]. 现代图书情报技术,2008(10):11-15.
[14] 祝犇. 数字信息长期保存中来源感知技术的研究[D]. 武汉:华中科技大学,2013.

[15] 余卉. 数字档案保存的来源感知系统设计与实现[D]. 武汉：华中科技大学,2012.
[16] 秦磊华,吴绪成,肖波. 数字资源长期保存的来源感知技术研究[J]. 现代教育技术,2013(12):102－106.
[17] Consultative Committee for Space Data Systems. Reference model for an open archival information system(OAIS):CCSDS 650.0－M－2[EB/OL].[2016－02－16]. http://public.ccsds.org/publications/archive/650x0m2.pdf.
[18] PREMIS data dictionary for preservation metadata, version 2.0[S/OL]. [2015－05－13]. http://www.loc.gov/standards/premis/v2/index.html.
[19] RAM S, LIU J. A semantic foundation for provenance management [J]. Journal on data semantics, 2012, 1(1):11－17.
[20] PREMIS Editorial Committee. PREMIS OWL ontology 2.2 now available[EB/OL]. [2015 － 03 － 31]. http://www.loc.gov/standards/premis/ontology－announcement.html.
[21] PROV-O: The PROV ontology [EB/OL]. [2015－05－17]. https://www.w3.org/TR/2013/REC－prov－o－20130430/.

作者贡献说明：
吴振新：提出研究思路,设计论文框架,修改论文,定稿；
李文燕：查询、收集相关论文资料,撰写论文初稿；
蒋世银：查询、收集相关论文资料。

Construction of a Provenance Management Framework for Long-term Preservation System Based on the Event-centric

Wu Zhenxin[1] Li Wenyan[1,2] Jiang Shiyin[1]
[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]KingChannels Co. Ltd, Beijing 100196

Abstract: [**Purpose/significance**] This paper studies constructing a provenance management framework for long-term preservation system, to ensure the authenticity, reliabilityand usability of data objects archived in the long-term preservation system by effectively managing provenance information. [**Method/process**] It makes the definition of provenance event based on the preservation lifestyle of digital objects, designs a provenance management framework according to the preservation flow of OAIS, and accomplishes the design of the management function model and provenance information model based on the event-centric. [**Result/conclusion**] The provenance management framework for long-term preservation system, which not only follows the relevant standards of long-term preservation but also takes into account the needs of practice, has a universality and feasibility for long-term preservation system, but still has a space for verification of effectiveness and practicability.

Keywords: provenance information provenance event long-term preservation lifestyle management OAIS PREMIS PROV ontology model